R-002

The MicrobesOnline Functional Genomics Website

K. Huang[1,4], M.N. Price[1,4], K. Keller[3], E. Alm[1,4], A.P. Arkin[1,2,3,4]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, [2]Howard Hughes Medical Institute, Berkeley, CA, USA, [3]Department of Bioengineering, University of California, Berkeley, USA

[4]Virtual Institute for Microbial Stress and Survival

Abstract

The Virtual Institute for Microbial Stress and Survival (VIMSS) has compiled a suite of software tools for use in the analysis of functional genomics experiments performed in prokaryotic systems. These results are integrated into the MicrobesOnline comparative genomics database (http://microbesonline.org/) to provide additional context and annotation of key responsive genes and pathways.

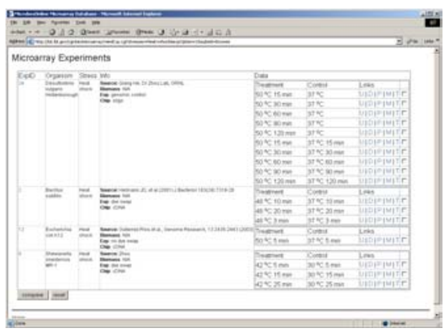
Methods: The current release of the Functional Genomics addition to the MicrobesOnline website focuses on in-depth analysis of transcriptomics data. Analysis includes data normalization for both direct ratio and genomic control experiments as well as significance testing.

Results: Once data is uploaded into the MicrobesOnline functional genomics database, a number of analytical tools are available. User scan browse a list of up- or down-regulated genes, or entire operons or start with an overview of (KEGG) metabolic maps colored according to expression trends. Agreement between changes in gene expression and predicted operon structures is provided as an independent biological validation. Overrepresented up- or down-regulated pathways and functions are available as a quick summary, and users can quickly compare the response of genes to different treatments in the same organism or compare the response of orthologous genes to similar treatments in different species. Currently, a large number of different experimental treatments covering a diverse set of species have been imported into the MicrobesOnline database.

Conclusions: The MicrobesOnline Functional Genomics database ties together analytical tools designed specifically for interpreting functional genomics studies in prokaryotes with a powerful set of tools for the comparative analysis of prokaryotic genomes.

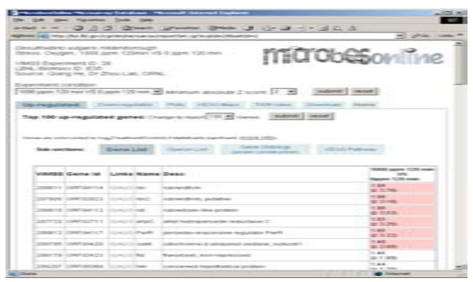
More than 35 experiments with >300 experimental conditions are in our microarray database. Most of the data in the database is currently unpublished and is generated by the VIMSS collaborators, therefore, the access is limited to our collaborators.

We plan to launch a separate public microarray website in the fall of 2005 with all the same website tools described here with previously published microarray data.



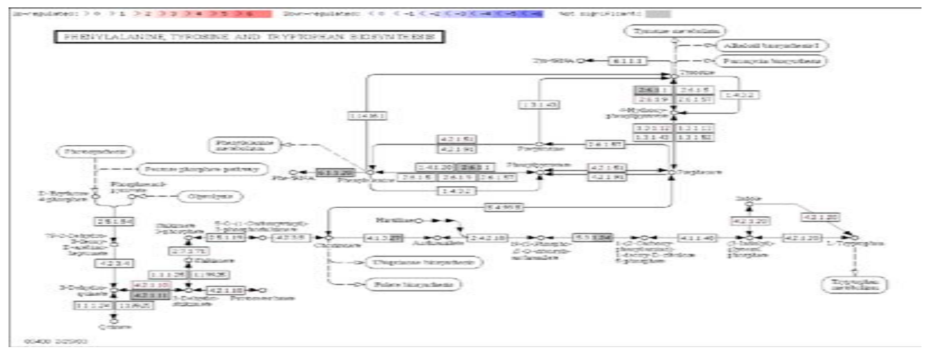
Gene List

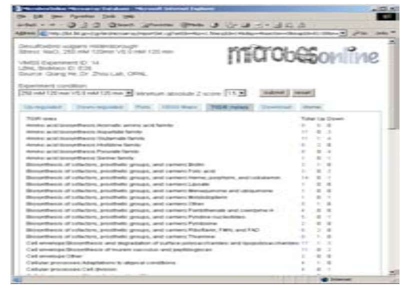
By default, top 100 up- or down-regulated genes are reported. If highlighted in color (red: up-regulated, blue: down-regulated), it indicates that the change of gene expression is statistically significant (i.e. >= the minimum absolute Z score, default |Z|=2). Options to change the minimum |Z| or the total number of genes to report are provided.



KEGG metabolic maps

KEGG metabolic maps are provided as a way to browse gene expression changes of the metabolic pathways under certain external environmental conditions. The change of gene expression of enzymes are indicated in colors (red: up-regulated, blue: down-regulated, grey:not significant). The option to change the minimum |Z|, which is used as a guideline of the significance of gene expression, is provided.
It is common that multiple genes share an EC number. In such cases, the rectangular box that represents each enzymatic activity is divided into vertical stripes, one for each gene with that EC number.



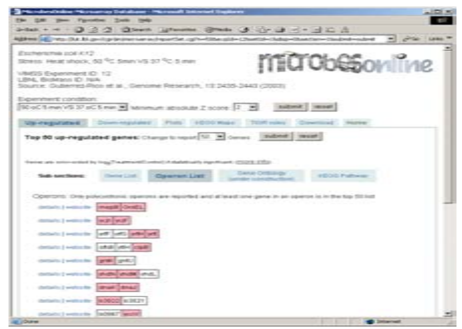


TIGR roles

Another way to browse the overall gene expression changes is using TIGR role function categories. TIGR roles are assigned through MicrobesOnline[2] automatic genome pipeline. Each row in the table shown in the figure above is a TIGR sub-role, where the total number of genes assigned to that role is listed, as well as the significantly up- and down-regulated genes according to the Z score cutoff. Hyperlinks are provided to view the gene expression changes and annotations in detail.

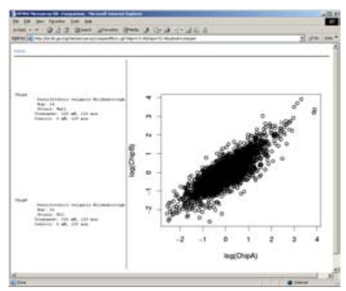
Operon List

By default, complete operon structures predicted by VIMSS[1] of the top 100 up- or down-regulated genes are reported. Again if highlighted in color, it indicates the change of gene expression is significant.



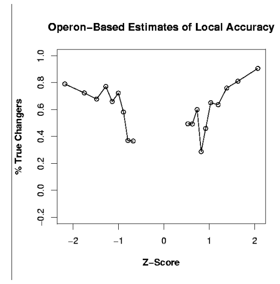
Scatter plot

The function of generating a scatter plot between two experiment conditions is provided. Users can select any two experiments, even two different species. When different species are chosen, only orthologous genes assigned through MicrobesOnline[2] automatic genome pipeline are used.



Operon-Based Estimates of Local Accuracy

Each point represents a group of 100 predicted significant changers with similar Z scores (the point shows the least significant Z value in the set). The estimated accuracy of each group of changers is derived by inspecting other genes in the same operons as these changers. Initially for random changers, 50% of these genes will move in the same direction, and for perfect changers, 100% of them will move in the same direction. Therefore, the total % of true changers can be estimated.



Members of these operons without a consistent signal across replicates (|Z|<0.5) are excluded. Even with perfect microarray data, the estimated accuracy will be somewhat less than 100% because of errors in the operon predictions.

Data Analysis

To estimate the differential gene expression between the control and treatment conditions, we use normalized log ratios. The log ratio is $\log_2$(treatment)- the $\log_2$(control). This log ratio is normalized using LOWESS on the difference vs. the sum of the log expression level. We have observed sector-based artifacts; therefore, the log ratio is further normalized by subtracting the median of all spots within each sector. Up to this point, the data processing is by spots instead of genes to allow sector-based normalization. Finally, we take the average of the spots for each gene to give a final normalized log ratio.

To assess the significance of the normalized log ratio, a Z-score to is calculated by the following equation:

$$Z = \frac{Log_2(Treatment\,/\,Control)}{\sqrt{0.25 + \sum \mathrm{var}iance}}$$

, where 0.25 is a pseudo-variance term.

For more info regarding the MicrobesOnline Functional Genomics Website, please contact gtlweb@vimss.lbl.gov

References

1. Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* 2005 Feb 8;33(3):880-92. Print 2005.

2. Alm EJ, Huang KH, Price MN, Koche R, Keller K, Dubchak I, Arkin AP. The MicrobesOnline Website for Comparative Genomics. *Genome Research, In Press.*

This work was part of the Virtual Institute for Microbial Stress and Survival supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC03-76SF00098 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

http://microbesonline.org